

# 周报

---

冯浩哲

2019.6.30

周报

4-6月工作进度总结与暑期工作安排

基于变分编码器的可视化图表的特征探索与标注工作

工作简介

工作进展

暑期工作安排

基于变分自编码器的半监督学习任务

工作简介

工作进展

模型细节

模型结果

暑期工作安排

本周工作汇报

下周工作安排

Reference

## 4-6月工作进度总结与暑期工作安排

---

从4月到6月以来，我主要进行了两个项目。一是基于变分编码器的可视化图表的特征探索与标注工作(基于今年1月份投稿IJCAI被拒后的工作继续改进，拟投稿今年9月AAAI)，二是基于变分编码器的半监督学习任务(拟投稿今年9月的ICLR)，工作简介与工作进展如下：

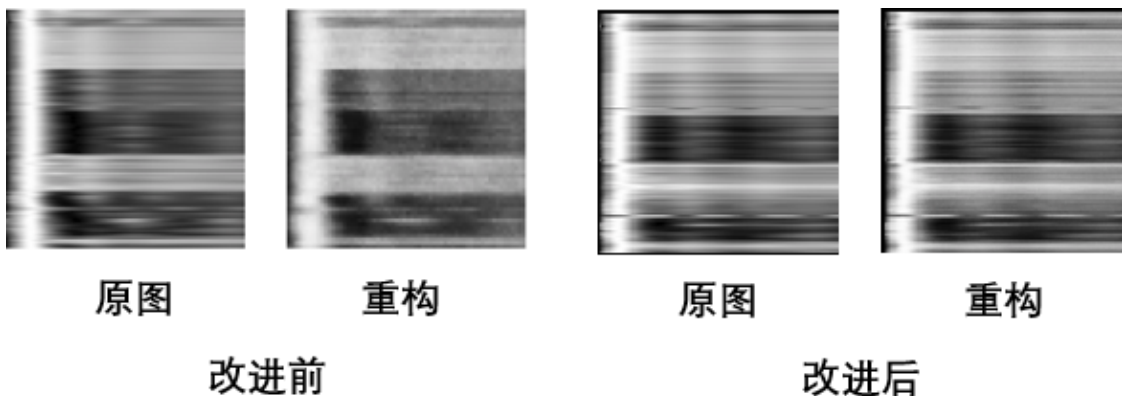
### 基于变分编码器的可视化图表的特征探索与标注工作

#### 工作简介

可视化图表能够帮助人们迅速发现并识别数据中的特征与规律。可是对于海量可视化图表，单靠人进行特征挖掘与标注则变得非常困难。我们基于变分编码器对可视化图表进行因子假设，利用无监督学习将可视化图表分解为低维空间中的独立因子，对每一张图表 $x$ 预测其因子的后验分布 $q(z|x)$ ，并构建从因子空间到原始空间的逆映射 $E_{q(z|x)}p(x|z)$ 。利用训练好的生成模型，我们可以基于分布距离度量(Wasserstein, kl-divergency)对图表进行聚类，类内代表性图表选取，图表间的连续变换(即如何在潜变量空间从一个可视化图表连续变换到另外一个可视化图表)以及每一个低维空间因子与图像对应的因子激活映射(Factor Activation Map, FAM)。利用这四种功能，我们可以对海量图表进行高效特征探索与标注。

#### 工作进展

1. 采用beta分布作为后验，解决了变分编码器在原电网像素图数据集上生成图像模糊的问题，现在变分编码器可以将电网数据映射到16维低维空间，并通过16维潜变量进行重构，重构结果与原图基本一致。



2. 原文在模型部分缺乏理论创新性, 我们提出了变分编码器因子的潜变量激活映射(**Factor Activation Map, FAM**), 从而对因子进行解释并作为我们的模型理论创新, 一些细节如下:

我们受[1],[2]的启发, 提出了该方法。[1],[2]文献针对全监督分类网络, 提出了利用第*i*类分类得分对特征图的梯度来对原图构造类别激活映射(Classification Activation Map)的方法:

$$\frac{\partial y_i}{\partial F_{i,j}} \quad (1)$$

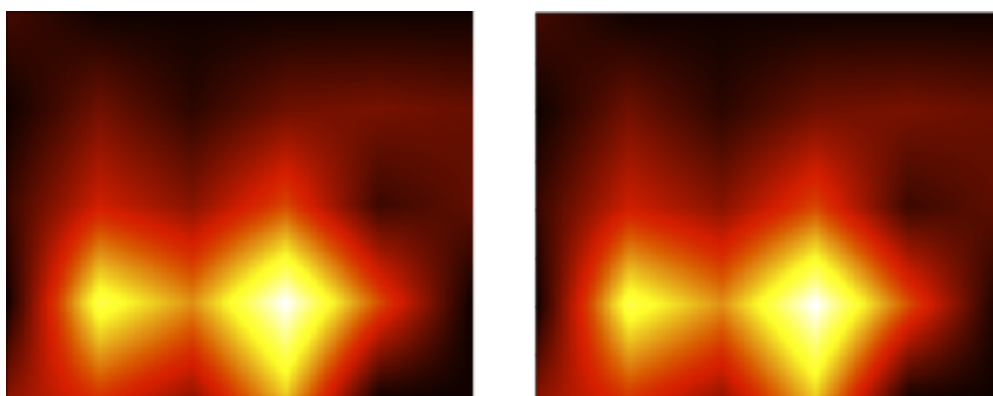
我们对(1)式进行变形, 得到了后验分布式符合变分编码器损失函数的形式

$$\begin{aligned} \frac{\partial y_c}{\partial F} &= y_c \frac{\partial \ln y_c}{\partial F} \\ &= y_c \frac{-\partial \sum_i l_i \ln \frac{l_i}{y_i}}{\partial F} \\ &= y_c - \frac{\partial \mathcal{D}_{KL}[p(y|x) \| q_\phi(y|x)]}{\partial F} \propto - \frac{\partial \mathcal{D}_{KL}[p(y|x) \| q_\phi(y|x)]}{\partial F} \end{aligned} \quad (2)$$

其中 $\mathcal{D}_{KL}[p(y|x) \| q_\phi(y|x)]$ 代表真实潜变量后验分布对预测潜变量后验分布的*kl-divergency*。但是在变分编码器的损失函数中, 我们并不知道真实后验 $p(y|x)$ 的形式, 因此我们的主要贡献在于提出了对(2)的估计

$$\frac{\partial \mathcal{D}_{KL}[q_\phi(z|x) \| p(z|x)]}{\partial F} \approx \frac{\partial (E_{q_\phi(z|x)} - \log p(x|z) + \mathcal{D}_{KL}[q_\phi(z|x) \| p(z)])}{\partial F} \quad (3)$$

右边是变分编码器可计算的损失函数, 对于估计(3)的有效性, 我们利用胸部X-ray数据集**CheXpert**在半监督VAE上的训练模型进行验证, 即对于有标签的数据, 同时计算(3)式左边与右边, 并给出相应因子的**FAM**比较如下



(3)式左

(3)式右

我们拟在1月份的文章[An Interactive Insight Identification and Annotation Framework for Power Grid Pixel Maps using DenseU-Hierarchical VAE](#)基础上进行改进，在7-8月完成以下三项任务，并投稿AAAI2020，我负责撰写实验设计以及论文的方法部分。

1. 在除了电网数据以外的几个公共数据集(现在找到了芝加哥城市犯罪热力图数据集)上实验我们的系统，以证明其可扩展性
2. 继续进行实验证明我们提出的变分编码器因子的潜变量激活映射(**Factor Activation Map, FAM**)的有效性，这是我们工作的重点，也是我们区别于1月份文章的主要创新点。

## 基于变分自编码器的半监督学习任务

### 工作简介

深度学习系统需要海量带标注数据进行训练，原始数据获取较为廉价，而获取准确的数据标注则成本很高。如何构造自监督任务(*self-supervised task*)，同时利用有标注数据与无标注数据进行训练是一个重要的问题。现在的深度半监督模型分为三个方向，一个是基于正则化目标函数的半监督分类[3] [4] [5]，一个是基于图模型(GNN)的半监督学习，还有一个是基于生成模型的半监督分类。基于生成模型的半监督分类模型有两种，基于GAN以及基于变分编码器(VAE)的模型。基于变分自编码器的半监督学习框架[6] [7] 将分类标签看成是一种离散潜变量 $c$ ，将分类任务看成是对离散潜变量的后验预测 $q_\phi(c|x)$ ，将有标注的部分看成是已知 $c$ 的真实后验分布 $p(c|x)$ ，在训练生成模型的过程中同时引导模型将后验预测 $q_\phi(c|x)$ 逼近真实后验分布 $p(c|x)$ 。

现有基于变分自编码器的半监督学习框架[6] [7]在SVHN与MNIST数据集上都有结果，分别发表在NIPS2016, NIPS2017，但是现有的框架有以下四个问题：

1. 模型基于高斯混合假设，对于有标注样本 $(x, y)$ 最大化其出现概率 $L_l(x, y)$ ，对于无标注样本 $(x)$ ，则从后验预测 $q_\phi(c|x)$ 中采样并最大化后验概率 $L_u(x) = \sum_{y \sim q_\phi(c|x)} q_\phi(y|x) L_l(x, y)$ ，这种方法对每个无标注样本都需要计算 $L_l(x, y)$   $C$ 次，计算量极大
2. 在训练过程中，模型需要2阶段层次结构，即先用经典VAE模型训练从 $x$ 到第一层潜变量 $z_1$ 的映射 $q(z_1|x)$ ，然后再从 $z_1$ 训练到第二层潜变量 $(z_2, c)$ 的映射 $q(z_2, c|z_1)$ ，最后结果严重依赖于 $z_1$ 习得表示的好坏，训练困难且结果不稳定
3. 现有框架的损失函数无法直接导出交叉熵，文献[6] [7]采用的方式是手动在损失函数上增加加权交叉熵，这破坏了整个变分编码器的理论推导体系，显得非常不自然
4. 对于多标签的情况，其损失函数变为 $L_u(x) = \sum_{y^1} \dots \sum_{y^k} q_\phi(y^1, \dots, y^k|x) L_l(x, y^1, \dots, y^k)$ ，计算复杂度为标签数的指数次 $O(k^c)$

我们现阶段主要提出了一种新框架，并对这4个问题进行了解决

### 工作进展

我们现阶段主要基于VAE提出了一种半监督学习的新框架，这个新框架可以解决以上四个问题，同时能够在SVHN与MNIST数据集上达到更好的分类结果。其细节与结果如下：

#### 模型细节

我们的模型基于潜变量边缘分布与条件分布的独立假设，将分类潜变量 $c$ 看作是模型的生成因子之一，损失函数最大化训练样本 $x$ 的出现概率 $\log p(x)$ ，模型的主要贡献在于对于有标注的场景，我们用Jensen不等式自然地导出了一个更紧的变分下界形式，同时这个形式可以自然地引出分类交叉熵，细节如(4), (5)所述：

$$\begin{aligned} -L_u(x) &= \log p(x) - \mathcal{D}_{KL}[q_\phi(z|x) \| p(z|x)] - \mathcal{D}_{KL}[q_\phi(c|x) \| p(c|x)] \\ &= E_{(z,c) \sim q_\phi(z,c|x)} \log p_\theta(x|z, c) - \mathcal{D}_{KL}[q_\phi(z|x) \| p(z)] - \mathcal{D}_{KL}[q_\phi(c|x) \| p(c)] \end{aligned} \quad (4)$$

$$\begin{aligned}
-L_l(x) &= \log p(x) \geq E_{z \sim q_\phi(z|x), c \sim p(c|x)} \log \frac{p(x, z, c)}{q_\phi(z|x)p(c|x)} \\
&= E_{z \sim q_\phi(z|x), c \sim p(c|x)} \log p(x|z, c) - \mathcal{D}_{KL}(q_\phi(z|x) \| p(z)) - \mathcal{D}_{KL}(p(c|x) \| q_\phi(c|x)) \\
&\quad + E_{c \sim p(c|x)} \log \frac{p(c)}{q_\phi(c|x)} \\
&\approx^{\text{when } q_\phi(c|x) \approx p(c|x)} E_{z \sim q_\phi(z|x), c \sim p(c|x)} \log p(x|z, c) - \mathcal{D}_{KL}(q_\phi(z|x) \| p(z)) - \mathcal{D}_{KL}(p(c|x) \| q_\phi(c|x)) \\
&\quad + E_{c \sim q_\phi(c|x)} \log \frac{p(c)}{q_\phi(c|x)} \\
&= E_{z \sim q_\phi(z|x), c \sim p(c|x)} \log p_\theta(x|z, c) - \mathcal{D}_{KL}[q_\phi(z|x) \| p(z)] - \mathcal{D}_{KL}[q_\phi(c|x) \| p(c)] \\
&\quad - \mathcal{D}_{KL}[p(c|x) \| q_\phi(c|x)]
\end{aligned} \tag{5}$$

这里 $-\mathcal{D}_{KL}[p(c|x) \| q_\phi(c|x)]$ 就是交叉熵，它的导出非常自然，同时减小了变分下界与真实似然分布 $\log p(x)$ 之间的 $margin$ ，给出了更紧的下界。

## 模型结果

我们的模型只需要一阶段训练就达到了比[6] [7]更好的结果，同时我们在胸部X-ray数据集**CheXpert**上实验了多标签(5标签)分类模型，也得到了较好的结果：

CheXpert Result				
Pathology	Baseline 100%	Baseline 50%	Ours 50%	Ours 10%
<b>Atelectasis</b>	0.808	0.803	<b>0.826</b>	<b>0.781</b>
<b>Cardiomegaly</b>	0.834	0.821	<b>0.855</b>	0.786
<b>Consolidation</b>	<b>0.904</b>	0.897	0.891	0.816
<b>Edema</b>	0.898	0.892	<b>0.9</b>	0.868
<b>Pleural Effusion</b>	0.921	0.904	0.904	<b>0.883</b>
MNIST Result				
	M1+TSVM[6]	M2[6]	M1+M2[6,7]	Ours
	11.82(±0.25)	11.97(±1.71)	3.33(±0.14)	<b>3.31(±0.19)</b>
SVHN Result				
	M1+KNN[6]	M1+TSVM[6]	M1+M2[6,7]	Ours
	65.63(±0.15)	54.33(±0.11)	36.02(±0.10)	<b>31.92(±0.14)</b>

## 暑期工作安排

当前对于原模型[6] [7]进行改进的工作已经全部做完，并拟写论文投稿**ICLR2020**。暑期的工作主要有两件：

### 1. 继续探索合理的多标签表达形式

我们当前对多标签分类问题仍然把它看作是多个独立的单标签分类，而没有考虑标签之间的相关性。这个假设不尽合理，我们需要用一些概率图改进这个假设

### 2. 采用正则化方法提高半监督结果[3] [4] [5]

当前基于正则化方法的半监督模型获得了最高的分类准确度，而基于VAE模型的准确度则没有那么多高。我们提出的新框架也可以应用正则化方法中的Mixup等策略[3],[5]，因此我们拟采用正则化方法改进结果，并将其作为我们的创新点

## 本周工作汇报

本周我主要集中精力在基于变分自编码器的半监督学习任务上，尝试采用正则化方法改进准确度。采用正则化方法做半监督的文章主要用了2类模型，一类是PreActResnet[8]，另外一类是WideResNet[9]，这两种模型都没有纳入Pytorch的官方模型，需要自己进行手动实验。

本周主要在 Cifar10, Cifar100, SVHN 上调参，并试图在分类任务上复现

PreActResnet18, PreActResnet34, PreActResnet101, PreActResnet152 以及 WideResNet-28-2, WideResNet-28-10, WideResNet-40-10 的结果。复现的过程中遇到了官方代码跑不通，论文中很多trick没有说明以及跑不出论文作者Claim的精度等障碍，但是基本还是复现了这些结果。

本周工作时长为60小时。

## 下周工作安排

---

下周主要是复现能应用于我们基于变分自编码器的半监督学习模型上的正则化方法[3],[5].

## Reference

---

1. Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2921-2929.
2. Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 618-626.
3. Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond empirical risk minimization[J]. arXiv preprint arXiv:1710.09412, 2017.
4. Verma V, Lamb A, Beckham C, et al. Manifold mixup: Encouraging meaningful on-manifold interpolation as a regularizer[J]. stat, 2018, 1050: 13.
5. Berthelot D, Carlini N, Goodfellow I, et al. Mixmatch: A holistic approach to semi-supervised learning[J]. arXiv preprint arXiv:1905.02249, 2019.
6. Kingma D P, Mohamed S, Rezende D J, et al. Semi-supervised learning with deep generative models[C]//Advances in neural information processing systems. 2014: 3581-3589.
7. Narayanaswamy S, Paige T B, Van de Meent J W, et al. Learning disentangled representations with semi-supervised deep generative models[C]//Advances in Neural Information Processing Systems. 2017: 5925-5935.
8. He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks[C]//European conference on computer vision. Springer, Cham, 2016: 630-645.
9. Zagoruyko S, Komodakis N. Wide residual networks[J]. arXiv preprint arXiv:1605.07146, 2016.